International Journal of Computer Networking,
Wireless and Mobile Communications (IJCNWMC)
ISSN (P): 2250-1568; ISSN (E): 2278-9448
Vol. 9, Issue 1, Jun 2019, 17-24

TRANS STELLAR
•Journal Publications •Research Consultancy

DEMYSTIFYING BIG DATA ANALYTICS TECHNIQUES

KOMALPREET KAUR[1], CHITENDER KAUR[2] & TARANDEEP KAUR BHATIA[3]

[1]*Research Scholar, Department of Computer Science & Engineering, Chandigarh Engineering College, Landran, Punjab, India*

[2]*Assistant Professor, Department of Computer Science & Engineering, Chandigarh Engineering College, Landran, Punjab, India*

[3]*Assistant Professor, Department of Computer Science & Engineering, Chitkara University, Punjab, India*

ABSTRACT

Big Data is relevant for massive amounts of data that do not have the potential to be managed by the traditional information processing techniques. This paper sights on introducing the term big data and illustrating various big data analytics techniques along with its benefits, bottlenecks and application areas that can be helpful in addressing the big data issues.

KEYWORDS: Big Data, Big Data Analytics, Sentiment Analysis & Neural Networks

Received: Jan 24, 2019; **Accepted:** Feb 14, 2019; **Published:** Mar 01, 2018; **Paper Id.:** IJCNWMCJUN20193

Original Article

INTRODUCTION

"BIG DATA" – a term representing vast and huge datasets with varying types of data and complex data structures having difficulties in storing, processing, analysing and retrieving information for generating the results [1]. Big data is of great significance to create profitable ventures in businesses and provides innumerable opportunities to achieve remarkable progress in various fields. In comparison to traditional datasets, big data mainly includes unstructured type of data that requires a thorough real-time analysis [2]. The procedure of analysing huge amounts of data to uncover the underlying patterns and connections is termed as big data analytics. These valuable information helps the organizations for gaining better insights and getting benefits over the competitive organizations [3]. For this reason, a precise analysis of big data is to be done.

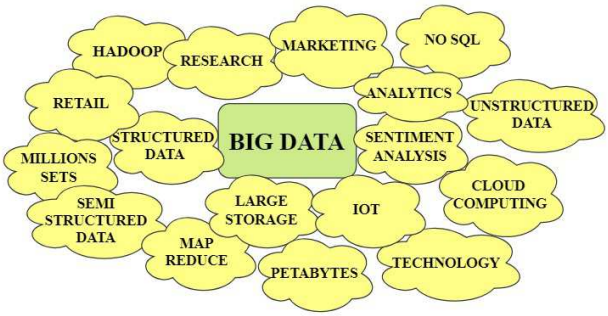

Figure 1: Big Data Concepts

LITERATURE REVIEW

Data analytics is a way of organizing big data. There are distinct patterns and inter-relationships within big data that helps data analytics to gain better insights regarding the characteristics of data. Thus, data analytics is the

most vital component of information technology.

Peter et al. [4] has discussed about various optimization methods for solving the optimization problems. These methods produce the solutions in a timely manner but consumes a great space in memory. Moreover, the authors have also mentioned the application areas for these methods such as biology, physics.

Chen et al. [5] has presented statistics by gathering the data and arranging it in such a manner that the underlying relationship between variables are explored. This has also been found that this technique is not appropriate for managing data in a better way. Many applications in the field of medicine and science have also been stated.

Hand et al. [6] has studied about data mining by performing thorough analysis on extensive datasets and also showed that this information has helped in predicting market trends in future to a great extent. The authors have also considered methods by which the data can be used for unethical purposes.

Pang et al. [7] has used various statistical methods to get a better understanding of machine learning concepts. The techniques have been discussed for making the computers capable of learning and acting along with utilization of resources in an efficient manner.

Potter et al. [8] has examined visualization approaches for generating diagrams, graphs and table to ease the analysis process of data. But, the authors have also shown that different users have different point of views which may produce inefficient results.

George et al. [9] has discussed about methods for comparing between two versions of an application. The authors have shown very effortless and easy analysis for performing A/B testing but have observed that this technique focuses on quantity only.

Lina et al. [10] has presented association learning methods for defining associations among the variables. The authors have also found that implementation of this technique is simple but not appropriate for large databases.

Lemon et al. [11] has described methods for making classifications regarding the categories to which observations belong. This technique turned out to be very complicated and unreliable for the authors. The authors has taken the example of predicting the type of soil in deserts.

Wilks et al. [12] has examined the work for cluster analysis by grouping the same objects in a cluster. The authors have shown the use of this technique for illustrating the hidden connections and also its contribution in the field of data mining.

Doan et al. [13] has discussed about crowd sourcing by assigning a task of journalism to the general public rather than any professional thus saving money greatly. But the authors have proved that because of the involvement of a large number of users, confidentiality is compromised.

Rosen et al. [14] has combined ensemble learning algorithms for obtaining better results and enhancing the stability of the system. The face recognition example has been thoroughly taken into account by the authors.

Whitley et al. [15] has described genetic algorithms for finding certain or approximate solutions for problems. But, huge amount of time has been taken for performing computations by the authors. The tasks related to pattern matching have been discussed.

Jain et al. [16] has explored interactions between the users and the system by manipulating natural languages such as text. The authors have used user- friendly techniques for faster processing and for performing sentiment analysis on huge dataset.

Liu et al. [17] has reviewed the way of processing information by neural networks that are inspired by the human brain. Apart from this, the authors have reported the difficulty in representing the problem to the network and dependency on hardware.

Fagan et al. [18] has focused on pattern recognition methods that requires human input. The deviations in the patterns have been taken into account by the authors. Moreover, the authors have undertaken example of stock markets for a better understanding of this technique.

Einav et al. [19] has analysed predictive modelling techniques for speculating the outcomes. The authors have explained the use of this method for implementing business strategies to achieve long term benefits but along with frequent updating in the systems.

DeFries et al. [20] has studied regression analysis by predicting the relationship between dependent variables and independent variables. The authors have used the fields of engineering, sales and economics to generate highly accurate results.

Agarwal et al. [21] has discussed about the methods for analysis of sentiments of users. Focus on improving the customer services and managing the crisis in an effective manner have been discussed by the authors. The authors have taken the examples of social networking sites.

Sandryhaila et al. [22] has manipulated the information in signals to convert into meaningful data. Authors have used cost- efficient methods for making the required changes in the program, but reported that synchronizing the systems for communication became a bottleneck.

Anselin et al. [23] has examined spatial analysis based on the conversion of data into useful information and identification of attributes in spatial data. The authors have shown in their work that requires of prior knowledge generated a lot of errors in the selection of sites.

Nadeau et al. [24] has described supervised learning techniques taking place under the supervision of a teacher. The authors have presented this technique as a highly specific one, but also stated that a lot of time was taken for performing the computations and training the network.

Table 1: Big Data Analytics Techniques

Sr. No	Name of the Technique	Description	Advantage	Disadvantage	Application Area
1.	Optimization methods[4]	Provides several strategies for computations and handling optimization problems.	Generates solutions for quantitative problems in minimal time.	Complexity in memory is high.	Speech, text processing, biology, physics
2.	Statistics[5]	Science that deals with collection and organization of data along with interpretation of results.	Explores the causal relationships among different variables.	Traditional statistical techniques are not appropriate for the well management of big data.	Scientific databases, Medical records.

3.	Data mining[6]	Process of analyzing large datasets for identifying the hidden patterns and relationships that are useful in solving a problem.	Helpful in predicting the future trends in the markets.	Data gathered through this technique for ethical motives can be fraudulently used.	Marketing, healthcare, medical science, banking, sales.
4.	Machine learning[7]	Science that uses statistical methods for providing ability to act and learn to the computers without programming explicitly.	Capability of handling heterogeneous data and utilizing the resources efficiently.	Acquisition is the main challenge associated with this technique. Data need to be thoroughly analyzed and processed before serving it as input to the given algorithm.	Extracting data from social networking websites, discovery of drugs, diagnosis in medical science, speech recognition.
5.	Visualization approaches[8]	Used in the creation of images, tables and generating diagrams, graphs to get better insights into data.	Provides quick access to business activities for better understanding of future trends.	Different users have different perceptions and this sometimes results in inaccurate results.	Retail marketing, cognition, apprehension.
6.	A/B Testing[9]	Method of drawing comparison between two versions of an application or a web page to conclude which performs better. Also known as split testing.	Provides effortless analysis and is multifunctional.	Speed is slow and is cost – ineffective. Moreover, focuses only on quantity and not quality.	Marketing, statistics, data analytics.
7.	Association rule learning[10]	Determines the relationships, that is, associations among different variables existing in large databases.	Scanning the repeated database is discarded, implementation is simple.	Suitable for only small databases, consumption of memory is large.	Market basket analysis dealing with the behaviors of customer purchases.
8.	Classification tree analysis[11]	Method of determining the category to which a new observation belongs to.	Easy to use, transparent and specific.	Expensive, complicated and unreliable.	Predicting the type of soil in the desert, making student profiles who opt for online courses.
9.	Cluster analysis [12]	Statistical tool for classification of objects into clusters, that is, groups in such a manner that all the objects that are similar to each other belong to the same group.	As it groups the similar objects together, therefore this technique is useful in identifying the patterns between the objects.	Difficulty in determining the number of clusters that are needed for analysis.	Exploratory research, data mining.
10.	Crowdsourcing[13]	A task, obtained from outside, is assigned to general public rather	Cost- effective, diversification in experienced	Confidentiality is at risk as a large group is involved in testing.	Astronomy, journalism, schedule of

		than any professional or an organization.	because of the involvement of large number of users.	Moreover, time and language barriers can make the communication between testers a bit tedious work.	events.
11.	Ensemble learning [14]	Combines different types of algorithms that are devised for learning so to enhance the stability of the model.	This technique can be effectively parallelized, thus enhancing the system performance.	In case the ensemble method chosen, is not appropriate according to the setting, then performance will decline automatically.	Detecting frauds, system security, face recognition.
12.	Genetic algorithms [15]	Search techniques applied in soft computing to determine the certain or approximate solutions for search problems	In case of hybrid applications, it can form blocks flexibly. Moreover, it is simple to understand.	Slower than other big data techniques as it takes more time for performing computations.	Training neural networks, performing tasks related to image processing like pattern matching.
13.	Natural Language Processing[16]	Deals with manipulating natural languages such as speech, text automatically by using software. It is basically concerned with system and human interactions.	User – friendly, processing is fast, infers solutions that are not generated before.	Precision and machine translation are major limitations of natural language processing.	Sentiment analysis, classifying the text, summarizing content, question answering.
14.	Neural networks [17]	Information processing technique influenced by the way information is processed by brain.	Fault – tolerant, can work with incomplete information.	Unpredictable behavior of network makes it difficult to represent problem to network, dependence on hardware.	Face recognition, speech recognition, and character recognition.
15.	Pattern recognition [18]	Section of machine learning that focuses on the recognition of patterns or deviations in a particular scenario.	Transparency is the merit of pattern recognition. Also, implemented manually with a great ease.	Human input is required, faces difficulty in case of complex datasets.	Forecasting stock markets, categorization of rocks, analyzing and controlling traffic.
16.	Predictive modelling[19]	Statistical technique for predicting future behavior. This technique uses probability and data mining to speculate outcomes.	Useful for business strategies in terms of taking decisions related to launching a new product in the market.	Requires updating of the model as the customer behavior varies with time.	Online marketing, advertising, weather forecasting, meteorology.

17.	Regression analysis[20]	Models the relationship between predictor variables and response variable. It describes how independent variables influences the dependent variables.	Results are highly accurate, uses multiple variables for gaining better insights.	Procedure for analysis and computations is complex and time consuming, not applicable for qualitative phenomena such as intelligence, honesty.	Engineering, economics, business sales.
18.	Sentiment analysis [21]	Method of determining the sentiments, that is, view or opinion of a person, writer or speaker on a particular topic.	Enhanced customer service, crisis are managed effectively.	Details like hilarity, anger, sarcasm are hardly noticeable using sentiment analysis.	Monitoring social media, exploiting unstructured data in business.
19.	Signal processing [22]	Manipulates the information in signals to simplify the speech recognition. It retrieves the information from speech and convert it into meaningful words.	Cost- efficient, processing operations can be changed by making changes in the program.	Proper synchronization of communication system becomes a major issue of this technique.	Seismology, biomedical engineering, control systems, audio processing.
20.	Spatial analysis [23]	Process of converting raw data into meaningful information. It is based on the identification of the location and attributes of spatial data.	Enhanced customer service, improved productivity, more efficient.	Requires prior knowledge and may generate a large number of errors.	Selection of sites, Visualization approaches.
21.	Supervised learning[24]	In supervised learning there is a teacher, under whose supervision, learning takes place. The input is given to network and corresponding output is generated which is then compared with desired output.	Extremely accurate as it is very specific in defining the variables.	A lot of computation time is needed for training the network.	Pattern, speech and handwriting recognition, database marketing.

CONCLUSIONS

The entrance into the age of big data has already taken place, which is the upcoming frontier in terms of innovation and productivity. This review paper provides a view of big data, including big data concepts. The big data analytics techniques are explored for solving the big data problems. Although some of the techniques are still not well established, but in the upcoming time these techniques will definitely face great advancements.

FUTURE SCOPE

The inception of big data unfolds incredible opportunities. During the IT period, technology was the major concern which leads to the generation of data. In the big data period, the worth of data will lead to the advancement of technologies in the near future. Big data will not just influence socially and economically in a positive way, but it will also impact everyone's way of thinking for one's benefit. Big data analytics will integrate into prospering market, thus providing boundless opportunities to the organizations and researchers. Moreover, with advanced technologies like Internet of Things (IOT), machine leaning, the future of big data is going to be unshakeable.

REFERENCES

1. *Michael, Katina, and Keith W. Miller. "Big data: New opportunities and new challenges [guest editors' introduction]." Computer 46, no. 6 (2013): 22-24.*
2. *Boyd, Danah, and Kate Crawford. "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon." Information, communication & society 15, no. 5 (2012): 662-679.*
3. *Gandomi, Amir, and Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analytics." International Journal of Information Management 35, no. 2 (2015): 137-144.*
4. *Richtárik, Peter, and Martin Takáč. "Parallel coordinate descent methods for big data optimization." Mathematical Programming 156, no. 1-2 (2016): 433-484.*
5. *Chen, James J., Eric Evan Chen, Weizhong Zhao, and Wen Zou. "Statistics in Big Data." Vol 53, no. 3 (2015): 186-202.*
6. *Hand, David J. "Principles of data mining." Drug safety 30, no. 7 (2007): 621-622.*
7. *Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pp. 79-86. Association for Computational Linguistics, 2002.*
8. *Chang, W. Y., & Chang, P. C. (2017). Application of Radial Basis Function Neural Network, to Estimate the State of Health for LFP Battery.*
9. *Potter, Kristin, Paul Rosen, and Chris R. Johnson. "From quantification to visualization: A taxonomy of uncertainty visualization approaches." In Uncertainty Quantification in Scientific Computing, pp. 226-249. Springer, Berlin, Heidelberg, 2012.*
10. *George, Gerard, Martine R. Haas, and Alex Pentland. "Big data and management." (2014): 321-326.*
11. *Lina, Lu, Chen Yaping, Yang Maishun, and Wei Hengyi. "Algorithm Optimization of Mining Association Rules [J]." COMPUTER ENGINEERING AND APPLICATIONS 8 (2000): 031.*
12. *Lemon, Stephenie C., Jason Roy, Melissa A. Clark, Peter D. Friedmann, and William Rakowski. "Classification and regression tree analysis in public health: methodological review and comparison with logistic regression." Annals of behavioral medicine 26, no. 3 (2003): 172-181.*
13. *Wilks, Daniel S. "Cluster analysis." In International geophysics, vol. 100, pp. 603-616. Academic press, 2011.*
14. *Doan, Anhai, Raghu Ramakrishnan, and Alon Y. Halevy. "Crowdsourcing systems on the world-wide web." Communications of the ACM 54, no. 4 (2011): 86-96.*
15. *Mengistu, A. D., & Alemayehu, D. M. (2016). Robot for visual object tracking based on artificial neural network. International Journal of Robotics Research and Development (IJRRD), 6(1), 1-6.*
16. *Rosen, Bruce E. "Ensemble learning using decorrelated neural networks." Connection science 8, no. 3-4 (1996): 373-384.*
17. *Whitley, Darrell. "A genetic algorithm tutorial." Statistics and computing 4, no. 2 (1994): 65-85.*
18. *Jain, Aditya, Gandhar Kulkarni, and Vraj Shah. "Natural language processing." International Journal of Computer Sciences and Engineering 6, no. 1 (2018).*
19. *Liu, Weibo, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E. Alsaadi. "A survey of deep neural network architectures and their applications." Neurocomputing 234 (2017): 11-26.*

20. *Fagan, Joseph F. "The origins of facial pattern recognition." In Psychological development from infancy, pp. 83-113. Routledge, 2017.*

21. *Einav, Liran, Amy Finkelstein, Sendhil Mullainathan, and Ziad Obermeyer. "Predictive modeling of US health care spending in late life." Science 360, no. 6396 (2018): 1462-1465.*

22. *Sundaramurthy, A., & Chitra, V. (2016). Big data gathering in wireless sensor network using hybrid dynamic energy routing protocol. BEST: International Journal of Management, Information Technology and Engineering (BEST: IJMITE), 4(4), 59-68.*

23. *DeFries, John C., and David W. Fulker. "Multiple regression analysis of twin data." Behavior genetics 15, no. 5 (1985): 467-473.*

24. *Agarwal, Apoorv, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. "Sentiment analysis of twitter data." In Proceedings of the workshop on languages in social media, pp. 30-38. Association for Computational Linguistics, 2011.*

25. *Sandryhaila, Aliaksei, and Jose MF Moura. "Big data analysis with signal processing on graphs: Representation and processing of massive data sets with irregular structure." IEEE Signal Processing Magazine 31, no. 5 (2014): 80-90.*

26. *Anselin, Luc, Ibnu Syabri, and Youngihn Kho. "GeoDa: an introduction to spatial data analysis." Geographical analysis38, no. 1 (2006): 5-22.*

27. *Nadeau, David, and Peter D. Turney. "A supervised learning approach to acronym identification." In Conference of the Canadian Society for Computational Studies of Intelligence, pp. 319-329. Springer, Berlin, Heidelberg, 2005.*